

How PBS Schedules Jobs

This article provides a simplified explanation of the PBS scheduling policy on Pleiades, Electra, and Endeavour. According to the current policy, jobs are sorted in the following order:

1. Current processor use (by mission directorate)
2. Job priority
3. Queue priority
4. Job size (wide jobs first)

However, job sorting is adjusted frequently in response to varying demands and workloads. In each scheduling cycle, PBS examines the jobs in sorted order, starting a job if it can. If the job cannot be started immediately, it is either rescheduled or simply bypassed for this cycle.

There are numerous reasons that the scheduler may not start a job, including:

- The queue has reached its maximum run limit
- Your job is waiting for resources
- Your mission share has run out
- The system is going into dedicated time
- Scheduling is turned off
- Your job has been placed on hold
- Your home filesystem or default /nobackup filesystem is down
- Your hard quota limit for disk space has been exceeded

For more information about each item in this list, see [Common Reasons Why Jobs Won't Start](#).

Note that a high-priority job might be blocked by one of the limits in this list, while a lower priority job from a different user or requesting fewer resources might not be blocked.

If your job is waiting in the queue, you can run the `qstat` command as follows to obtain information that can indicate why it has not started running.

```
pfe21% qstat -s job_id
or
pfe21% qstat -f job_id | grep -i comment
```

On Pleiades, output from the `qs` command shows the amount of resources used and borrowed by each mission directorate, and the resources each mission is waiting for, by processor type:

```
pfe21% /u/scicon/tools/bin/qs
```

The following command provides the order of jobs that PBS schedules to start at the current scheduling cycle. It also provides information regarding processor type(s), mission, and job priority:

```
pfe21% qstat -W o=+model,mission,pri -i
```

Note: To prevent jobs from languishing in the queues for an indefinite time, PBS reserves resources for the top *N* jobs (between 6 and 12), and doesn't allow lower-priority jobs to start if they would delay the start time of a higher-priority job. This is known as [backfilling](#).

PBS Sorting Order

Mission Shares

Each NASA mission directorate is allocated a certain percentage of the processors on Pleiades and Electra. (See [Mission Shares Policy on Pleiades](#).) A job cannot start if that action would cause the mission to exceed its share, unless another mission is using less than its share and has no jobs waiting. In this case, the high-use mission can "borrow" processors from the lower-use mission for up to a specified time (currently, max_borrow is 4 hours).

So, if the job itself needs less than max_borrow hours to run, or if a sufficient number of other jobs from the high-use mission will finish within max_borrow hours to get back under its mission share, then the job can borrow processors.

When jobs are sorted, jobs from missions using less of their share are picked before jobs from missions using more of their share.

Job Priority

Job priority has three components. First is the native priority (the `-p` parameter to `qsub` or `qalter`). Added to that is the queue priority. If the native priority is 0, then a further adjustment is made based on how long the job has been waiting for resources. Waiting jobs get a "boost" of up to 20 priority points, depending on how long they have been waiting and which queue they are in.

This treatment is modified for queues assigned to the Human Exploration and Operations Mission Directorate (HEOMD). For those queues, job priority is set by a separate set of policies controlled by HEOMD management.

Queue priority

Some queues are given higher or lower priorities than the default (run `qstat -Q` to get current values). Note that because the mission share is the most significant sort criterion, job and queue priorities have little effect mission-to-mission.

Job Size

Jobs asking for more nodes are favored over jobs asking for fewer. The reasoning is that, while it is easier for narrow jobs to fill in gaps in the schedule, wide jobs need help collecting enough CPUs to start.

Backfilling

The policy described above could result in a large, high-priority job being blocked indefinitely by a steady stream of smaller, low-priority jobs. To prevent jobs from languishing in the queues for an indefinite time, PBS reserves resources for the top N jobs (N is between 6 and 12), and doesn't allow lower priority jobs start if they would delay the start time of one of the top job ("backfilling"). Additional details are given below.

As mentioned above, when PBS cannot start a job immediately, if it is one of the first N such jobs, PBS sets aside resources for the job before examining other jobs. That is, PBS looks at the currently running jobs to see when they will finish (using the wall-time estimates). From those finish times, PBS decides when enough resources (such as CPUs, memory, mission share, and job limits) will become available to run the top job.

PBS then creates a virtual reservation for those resources at that time. Now, when PBS looks at other jobs to see if they can start immediately, it also checks whether starting the job would collide with one of these reservations. Only if there are no collisions will PBS start the lower priority jobs.

Article ID: 179

Last updated: 16 Feb, 2018

Revision: 23

Running Jobs with PBS -> How PBS Schedules Jobs

<https://www.nas.nasa.gov/hecc/support/kb/entry/179/>